Description of Gene Expression Ratio Data File for BRCA Experiments

1. Contained in the downloadable data file are the gene expression ratios from 21 microarray experiments. The format of the file is,
    a. Tab-delimited text file
    b. The first row provides Patient ID, cited in the NEJM paper, for each experiments {1, 2, …, 21}.
    c. The second row provides mutation classification for each experiment, {BRCA1, BRCA2, Sporadic}.
    d. The third row provides experiment ID, {s1996, s1822, etc}.

    e. The first column is the microtiter plate ID where each clone physically locates.
    f. The second column is the IMAGE Clone ID, which can be used to perform database lookup.
    g. The third column is the Clone Title.
    h. The 4th to 24th columns contain gene expression ratio for each gene in each experiment.
2. Gene expression ratios included in the data file were derived from the fluorescent intensity (proportional to the gene expression level) from a tumor sample (BRCA1, BRCA2, or Sporadic) divided by the fluorescent intensity from a common reference sample (MCF-10A cell line). The common reference sample is used for all 21 microarray experiments. Therefore, the ratio may take value from 0 to infinity. (There is no negative value in the data table.)
3. We select these genes based on following criterion,
    a. Average fluorescent intensity (level of expression) of more than 2,500 (gray level) across all 21 samples,
    b. Average spot area of more than 40 pixels across all 21 samples, and
    c. No more than one sample in which the spot area is zero pixel.

    There are total of 3226 genes satisfy these requirements and thus included in the downloadable data file.
4. Ratios, included in the downloadable data file, for each experiments were normalized (or calibrated) such that the majority of the gene expression ratios from a pre-selected internal control gene set was around 1.0.
5. In most of the data analysis methods cited in the paper, we performed a logarithm-transform to convert the ratio data in order to achieve the symmetric property from over-expression to under-expression range. These methods include (but not limit to) MDS, weighted gene analysis, Class Prediction, *F*-test and *t*-test, and InfoScore method. We provide the normalized ratio data (NO log-transformation!) in the downloadable file such that some other data preprocessing methods may be attempted.

Prepared by Yidong Chen
CGB/NHGRI/NIH
May 16, 2001